

Functional Horseshoe Priors for Subspace Shrinkage

Minsuk Shin, Anirban Bhattacharya, and Valen E. Johnson

Texas A&M University, College Station, U.S.A.

Abstract

We introduce a new shrinkage prior on function spaces, the functional horseshoe prior, that encourages shrinkage towards parametric classes of functions. Unlike existing shrinkage priors for parametric models, the shrinkage acts on the shape of the function rather than sparsity of model parameters. We compare its performance with classical nonparametric estimators and a number of penalized likelihood approaches, and we show that the proposed procedure outperforms the competitors in the considered simulations and real examples. The proposed prior also provides a natural penalization interpretation, and casts light on a new class of penalized likelihood methods for function estimation. We theoretically exhibit the efficacy of the proposed approach by showing an adaptive posterior concentration property.

1 Introduction

Since the seminal work of James and Stein (1961), shrinkage estimation has been immensely successful in various statistical disciplines and continues to enjoy widespread attention. Many shrinkage estimators have a natural Bayesian flavor. For example, one obtains the ridge regression estimator as the posterior mean arising from an isotropic Gaussian prior on the vector of regression coefficients (Hoerl and Kennard, 1970; Jeffreys, 1961). Along similar lines, an empirical Bayes interpretation of the (positive part) James–Stein estimator can be obtained (Efron and Morris, 1973). Such connections have been extended to the semiparametric regression context, with applications to smoothing splines and penalized splines (Ruppert et al., 2003; Wahba, 1990). Over the past decade and a half, a number of second-generation shrinkage priors have appeared in the literature in relation to high-dimensional sparse estimation. Such priors can be almost exclusively expressed as global-local scale mixtures of Gaussians (Polson and Scott, 2010); examples include the relevance vector machine (Tipping, 2001), normal/Jeffrey’s prior (Bae and Mallick, 2004), the Bayesian lasso (Hans, 2009; Park and Casella, 2008), the horseshoe (Carvalho et al., 2010), normal/gamma and normal/inverse-Gaussian priors (Caron and Doucet, 2008; Griffin and Brown, 2010), generalized double Pareto priors (Armagan et al., 2013) and Dirichlet–Laplace priors (Bhattacharya et al., 2015). These priors typically have a large spike near zero with heavy tails, thereby providing an approximation to the operating characteristics of sparsity inducing discrete mixture priors (George and McCulloch, 1997; Johnson and Rossell, 2012). For more on connections between Bayesian model averaging and shrinkage, refer to Polson and Scott (2010).

A key distinction between ridge-type shrinkage priors and the global-local priors is that while ridge-type priors typically shrink towards a fixed point, most commonly the origin, the global-local priors provide shrinkage towards the union of subspaces consisting of sparse vectors, with the amount of sparsity controlled by certain hyperparameters (Bhattacharya et al., 2015). In this article, we further enlarge the scope of shrinkage to present a class of functional shrinkage priors, namely the functional horseshoe priors (fHS), that facilitate shrinkage towards pre-specified subspaces. The shrinkage factor (defined in Section 3) is assigned a

Beta(a, b) prior with $a, b < 1$, which has the shape of a horseshoe (Carvalho et al., 2010). However, while the horseshoe prior of Carvalho et al. (2010) shrinks towards sparse vectors, the proposed functional horseshoe prior shrinks functions towards arbitrary subspaces.

As a preliminary example, consider a nonparametric regression model with unknown regression function $f : \mathcal{X} \rightarrow \mathbb{R}$ given by

$$Y = F + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \mathbf{I}_n), \quad (1)$$

where $Y = (y_1, \dots, y_n)^\top$, and $F = (f(x_1), \dots, f(x_n))^\top = \mathbb{E}(Y \mid \mathbf{x})$, with the covariates $x_i \in \mathcal{X} \subset \mathbb{R}$.

In (1), we can either make parametric assumptions (e.g., linear or quadratic dependence on x) regarding the shape of f , or model it nonparametrically using splines, wavelets, Gaussian processes, etc. Although one can examine a scatter plot or perform a goodness of fit test to ascertain the validity of a linear or quadratic model in (1), such an exercise is only feasible in relatively simple settings. In relatively complex and/or high dimensional problems, there is clearly a need for an automatic data-driven procedure to adapt between models of varying complexity. With this motivation, we propose the functional horseshoe prior that encourages shrinkage towards a parametric class of models embedded inside a larger semiparametric model, as long as a suitable projection operator can be defined. For example, in (1), f will be shrunk towards a linear or quadratic function if such parametric assumptions are supported by the data, and will remain unshrunk otherwise. As noted already, our approach is not limited to the univariate regression context and can be extended to the varying coefficient model (Hastie and Tibshirani, 1993), density estimation via log-spline models (Kooperberg and Stone, 1991), and additive models (Hastie and Tibshirani, 1986), among others; further details are provided in Section 5. In the additive regression context, the proposed approach is highly competitive to state-of-the-art procedures such as the *Sparse Additive Model* (SpAM) of Ravikumar et al. (2009) and the *High-dimensional Generalized Additive Model* (HGAM) by Meier et al. (2009).

We provide theoretical support to the method by showing an adaptive property of the approach in the context of (1). Specifically, we show that the posterior contracts (Ghosal et al., 2000) at the parametric rate if the true function belongs to the pre-designated subspace, and contracts at the optimal rate for α -smooth functions otherwise. In other words, our approach adapts to the parametric shape of the unknown function while allowing deviations from the parametric shape in a nonparametric fashion.

2 Preliminaries

We begin by introducing some notation. For $\alpha > 0$, let $\lfloor \alpha \rfloor$ denote the largest integer smaller than or equal to α and $\lceil \alpha \rceil$ denote the smallest integer larger than or equal to α . Let $C^\alpha[0, 1]$ denote the Hölder class of α smooth functions on $[0, 1]$ that have continuously differentiable derivatives up to order $\lfloor \alpha \rfloor$, with the $\lfloor \alpha \rfloor$ th order derivative being Lipschitz continuous of order $\alpha - \lfloor \alpha \rfloor$. For a vector $x \in \mathbb{R}^d$, let $\|x\|$ denote its Euclidean norm. For a function $g : [0, 1] \rightarrow \mathbb{R}$ and points $x_1, \dots, x_n \in [0, 1]$, let $\|g\|_{2,n}^2 = n^{-1} \sum_{i=1}^n g^2(x_i)$; we shall refer to $\|\cdot\|_{2,n}$ as the empirical L_2 norm. For an $m \times d$ matrix A with $m > d$ and $\text{rk}(A) = d$, let $\mathfrak{L}(A) = \{A\beta : \beta \in \mathbb{R}^d\}$ denote the column space of A , which is a d -dimensional subspace of \mathbb{R}^m . Let $Q_A = A(A^\top A)^{-1}A^\top$ denote the projection matrix on $\mathfrak{L}(A)$.

3 The functional horseshoe prior

In the nonparametric regression model in (1), we model the unknown function f as spanned by a set of pre-specified basis functions $\{\phi_j\}_{1 \leq j \leq k_n}$ as follows:

$$f(x) = \sum_{j=1}^{k_n} \beta_j \phi_j(x). \quad (2)$$

We shall work with the B-spline basis in the sequel, though the methodology generalizes to a larger class of basis functions. The B-splines basis functions can be constructed in a recursive way. Let the positive integer q denote the degree of the B-spline basis functions satisfying $k_n > q + 1$. Define a sequence of knots $0 = t_0 < t_1 < \dots < t_{k_n-q} = 1$. In addition, define q knots $t_{-q} = \dots = t_{-1} = t_0$ and another set of q knots $t_{k_n-q} = \dots = t_{k_n}$. As in De Boor (2001), the B-spline basis functions are defined as

$$\begin{aligned} \phi_{j,1}(x) &= \begin{cases} 1, & t_j \leq x < t_{j+1}, \\ 0, & \text{otherwise,} \end{cases} \\ \phi_{j,q+1}(x) &= \frac{x - t_j}{t_{j+q} - t_j} \phi_{j,q}(x) + \frac{t_{j+q+1} - x}{t_{j+q+1} - t_{j+1}} \phi_{j+1,q}(x), \end{aligned}$$

for $j = -q, \dots, k_n - q - 1$. We reindex $j = -q, \dots, k_n - q - 1$ to $j = 1, \dots, k_n$ and the number of basis functions is k_n .

Letting $\beta = (\beta_1, \dots, \beta_{k_n})^T$ denote the vector of basis coefficients and $\Phi = \{\phi_j(X_i)\}_{1 \leq i \leq n, 1 \leq j \leq k_n}$ denote the $n \times k_n$ matrix of basis functions evaluated at the observed covariates, the model (1) can be expressed as

$$Y \mid \beta \sim N(\Phi\beta, \sigma^2 \mathbf{I}_n). \quad (3)$$

A standard choice for a prior on β is a g -prior $\beta \sim N(0, g(\Phi^T \Phi)^{-1})$ (Zellner, 1986). g -priors have been commonly used in linear models, since they incorporate the correlation structure of the covariates inside the prior variance. The posterior mean of β with a g -prior can be expressed as $\{1 - 1/(1 + g)\} \hat{\beta}$, where $\hat{\beta} = \mathbf{Q}_\Phi Y$ is the maximum likelihood estimate of β . Thus, the posterior mean shrinks the maximum likelihood estimator towards zero, with the amount of shrinkage controlled by the parameter g . Bontemps (2011) studied asymptotic properties of the resulting posterior by providing bounds on the total variation distance between the posterior distribution and a Gaussian distribution centered at the maximum likelihood estimator with the inverse Fisher information matrix as covariance. In his work, the g parameter was fixed *a priori* depending on the sample size n and the error variance σ^2 . His results in particular imply minimax optimal posterior convergence for α -smooth functions. Among related work, Ghosal and van der Vaart (2007) established minimax optimality with isotropic Gaussian priors on β .

Our goal is to define a broader class of shrinkage priors on β that facilitate shrinkage towards a *null subspace* that is fixed in advance, rather than shrinkage towards the origin or any other fixed *a priori* guess β_0 . For example, if we have *a priori* belief that the function is likely to attain a linear shape, then we would like to impose shrinkage towards the class of linear functions. In general, our methodology allows shrinkage towards any null subspace spanned by the columns of a null regressor matrix Φ_0 , with $d_0 = \text{rank}(\Phi_0)$ the dimension of the null space. For example in the linear case, we define the null space as $\mathfrak{L}(\Phi_0)$ with $\Phi_0 = \{\mathbf{1}, \mathbf{x}\} \in \mathbb{R}^{n \times 2}$, where $\mathbf{1}$ is a $n \times 1$ vector of ones and $d_0 = 2$. Shrinkage towards quadratic, or more generally polynomial, regression are achieved similarly.

With the above ingredients, we propose the functional horseshoe prior through the following hierarchical

specification:

$$\pi(\beta \mid \tau) \propto (\tau^2)^{-(k_n - d_0)/2} \exp \left\{ -\frac{1}{2\sigma^2\tau^2} \beta^\top \Phi^\top (I - Q_0) \Phi \beta \right\}, \quad (4)$$

$$\pi(\tau) \propto \frac{(\tau^2)^{b-1/2}}{(1 + \tau^2)^{(a+b)}} \mathbb{1}_{(0, \infty)}(\tau), \quad (5)$$

where $a, b > 0$ and recall that $Q_0 = \Phi_0(\Phi_0^\top \Phi_0)^{-1} \Phi_0^\top$ denotes the projection matrix of Φ_0 .

When $\Phi_0 = 0$, (4) is equivalent to a g -prior with $g = \tau^2$. The key additional feature in our proposed prior is to introduce the quantity $(I - Q_0)$ in the exponent, which enables shrinkage towards subspaces rather than single points. Although the proposed prior may be singular, it follows from the subsequent results that the joint posterior of (β, τ^2) is proper. Note that the prior on the scale parameter τ follows a half-Cauchy distribution when $a = b = 1/2$. Half-Cauchy priors have been recommended as a default prior choice for global scale parameters in the linear regression framework (Polson and Scott, 2012). Using the reparameterization $\omega = 1/(1 + \tau^2)$, the prior (5) can be interpreted as the prior induced on τ via a Beta(a, b) prior on ω . We shall work in the ω parameterization subsequently for reasons to be evident shortly.

Exploiting the conditional Gaussian specification, the conditional posterior of β is Gaussian,

$$\beta \mid Y, \omega \sim N(\tilde{\beta}_\omega, \tilde{\Sigma}_\omega), \quad (6)$$

where

$$\tilde{\beta}_\omega = \left(\Phi^\top \Phi + \frac{\omega}{1 - \omega} \Phi^\top (I - Q_0) \Phi \right)^{-1} \Phi^\top Y, \quad \tilde{\Sigma}_\omega = \sigma^2 \left(\Phi^\top \Phi + \frac{\omega}{1 - \omega} \Phi^\top (I - Q_0) \Phi \right)^{-1}. \quad (7)$$

We now state a lemma which delineates the role of ω as the parameter controlling the shrinkage.

Lemma 3.1. *Suppose that $\mathcal{L}(\Phi_0) \subsetneq \mathcal{L}(\Phi)$. Then,*

$$\mathbb{E}[\Phi \beta \mid Y, \omega] = \Phi \tilde{\beta}_\omega = (1 - \omega) Q_\Phi Y + \omega Q_0 Y,$$

where Q_Φ is the projection matrix of Φ .

The above lemma suggests that the conditional posterior mean of the regression function given ω is a convex combination of the classical B-spline estimator $Q_\Phi Y$ and the parametric estimator $Q_0 Y$. The parameter $\omega \in (0, 1)$ controls the shrinkage effect; the closer ω is to 1, the greater the shrinkage towards the parametric estimator. We learn the parameter ω from the data with a Beta(a, b) prior on ω . The hyperparameter $b < 1$ controls the amount of prior mass near one.

Figure 1 illustrates the connection between the choice of the hyperparameters a and b and the shrinkage behavior of the prior. The first and the second column in Figure 1, with a fixed at $1/2$ shows that the prior density of ω increasingly concentrates near 1 as b decreases from $1/2$ to 10^{-1} . The third column in Figure 1 depicts the prior probability that $\omega > 0.95$ and $\omega < 0.05$. Clearly, as b decreases, the amount of prior mass around one increases, which results in stronger shrinkage towards the parametric estimator. In particular, when $a = b = 1/2$, the resulting ‘‘horseshoe’’ prior density derives its name from the shape of the prior on ω (Carvalho et al., 2010).

When $\mathcal{L}(\Phi_0) \subsetneq \mathcal{L}(\Phi)$, we can orthogonally decompose $Q_\Phi = Q_1 + Q_0$, where the columns of Q_1 are orthogonal to Q_0 , i.e., $Q_1^\top Q_0 = 0$. To see this, since $\mathcal{L}(\Phi_0) \subsetneq \mathcal{L}(\Phi)$, we can use Gram-Schmidt orthogonalization to create $\tilde{\Phi} = [\Phi_0; \Phi_1]$ of the same dimension as Φ such that $\Phi_1^\top \Phi_0 = 0$ and $\mathcal{L}(\Phi) = \mathcal{L}(\tilde{\Phi})$. Then, we let Q_1 denote the projection matrix on $\mathcal{L}(\Phi_1)$. Simple algebra shows that

$$\begin{aligned} \pi(\omega \mid Y) &= \int \pi(\omega, \beta \mid Y) d\beta = \frac{\pi(\omega)}{m(Y)} \int f(Y \mid \beta, \omega) \pi(\beta \mid \omega) d\beta \\ &= \omega^{a+(k_n - d_0)/2 - 1} (1 - \omega)^{b-1} \exp\{-H_n \omega\} / m(Y), \end{aligned} \quad (8)$$

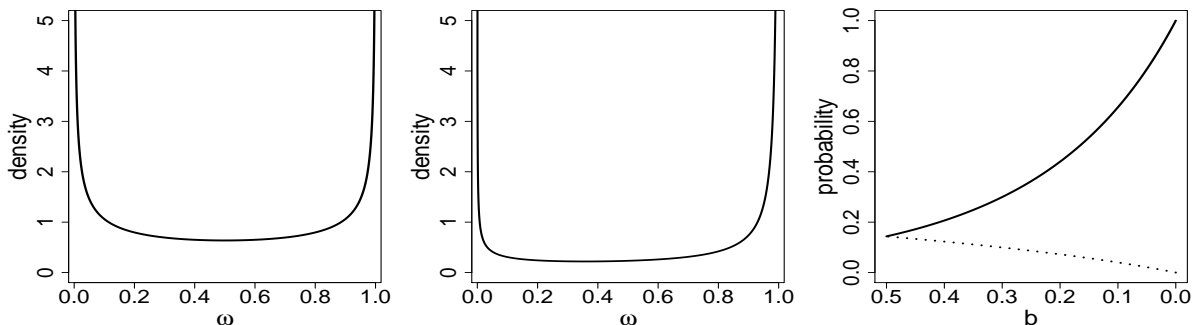


Figure 1: The first two columns illustrate the prior density function of ω with different hyperparameters (a, b) : $(1/2, 1/2)$ for the first column and $(1/2, 10^{-1})$ for the second column. The third column shows the prior probability that $\omega > 0.95$ (solid line) and $\omega < 0.05$ (dotted line) for varying b and a fixed $a = 1/2$.

where $H_n = Y^T Q_1 Y / (2\sigma^2)$ and $m(Y) = \int_0^1 \omega^{a+(k_n-d_0)/2-1} (1-\omega)^{b-1} \exp\{-H_n \omega\} d\omega$.

To investigate the asymptotic behavior of the implied posterior, it is crucial to find tight two-sided bounds on $m(Y)$, which is stated in Lemma 3.2.

Lemma 3.2. (Bounds on the normalizing constant) Let A_n and B_n be arbitrary sequences satisfying $A_n \rightarrow \infty$ as $n \rightarrow \infty$ and $B_n = O(1)$. Define $t_n = \int_0^1 \omega^{A_n-1} (1-\omega)^{B_n-1} \exp\{-H_n \omega\} d\omega$. Then,

$$\frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n+B_n)} \exp\{-H_n\} (1+Q_n^L) \leq t_n \leq \frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n+B_n)} \exp\{-H_n\} (1+Q_n^U),$$

where,

$$\begin{aligned} Q_n^U &= \frac{B_n}{A_n+B_n} \exp(H_n), \\ Q_n^L &= \frac{B_n H_n}{A_n+B_n} + \frac{D B_n (B_n+T_n)^{-A_n}}{(A_n+B_n)^{3/2}} \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \right)_+, \end{aligned}$$

where $T_n = \max\{A_n^2, 3\lceil H_n \rceil\}$ and D is some positive constant.

By setting $A_n = a + k_n/2$ and $B_n = b$, Lemma 3.2 suggests that the magnitude of the normalizing constant $m(Y)$ in (8) is determined by an interplay between the relative sizes of b and $\exp(H_n)$. When b is small enough to dominate $\exp(H_n)$, $m(Y) \approx \text{Be}(a + k_n/2, b) \exp(-H_n)$, where $\text{Be}(\cdot, \cdot)$ denotes the beta function. Otherwise, $m(Y) \approx \text{Be}(a + k_n/2, b)b$ ignoring polynomial terms. This asymptotic behavior of $m(Y)$ is the key ingredient to identify the posterior contraction rate of the functional horseshoe prior. We also note that the magnitude of a asymptotically does not affect the strength of shrinkage as long as a is a fixed constant, since the prior contribution ω^{a-1} is dominated by the likelihood contribution $\omega^{k_n/2}$.

3.1 Posterior concentration rate

We state a set of assumptions (Zhou et al. (1998), Claeskens et al. (2009)) that have been used in the literature to prove minimax optimality of B-spline estimators. Assume that the following conditions hold:

(A1). Let $u = \max_{1 \leq j \leq (k_n-1)} (t_{j+1} - t_j)$. There exists a constant $C > 0$, such that $u / \min_{1 \leq j \leq (k_n-1)} (t_{j+1} - t_j) \leq C$ and $u = o(k_n^{-1})$.

(A2). There exists some distribution function G with a positive continuous density such that

$$\sup_{x \in [0,1]} |G_n(x) - G(x)| = o(k_n^{-1}),$$

where G_n is the empirical distribution of the covariates $\{x_i\}_{1 \leq i \leq n}$, which are fixed by design.

Under **(A1)** and **(A2)**, Zhou et al. (1998) showed that the mean square error of the B-spline estimator $Q_\Phi Y$ achieves the minimax optimal rate. If the true function $f_0 \in C^\alpha[0, 1]$ is α -smooth and the number of basis functions $k_n \asymp n^{1/(2\alpha+1)}$, then Zhou et al. (1998) shows that

$$\mathbb{E}_0 \left[\|Q_\Phi Y - F_0\|_{2,n}^2 \right] = O \left(n^{-2\alpha/(1+2\alpha)} \right), \quad (9)$$

where $\mathbb{E}_0(\cdot)$ represents an expectation with respect to the true data generating distribution of Y .

We now state our main result on the posterior contraction rate of the functional horseshoe prior.

Theorem 3.3. *Consider the model (1) equipped with the functional horseshoe prior (4)-(5). Assume **(A1)** and **(A2)** hold and $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$. Further, assume that for some integer $\alpha \geq 1$, the true regression function $f_0 \in C^\alpha[0, 1]$ and the B-spline basis functions Φ are constructed with $k_n - \lfloor \alpha \rfloor$ knots and $\lfloor \alpha \rfloor - 1$ degree, where $k_n \asymp n^{1/(1+2\alpha)}$. Suppose that the prior hyperparameters a and b in (5) satisfy $a \in (\delta, 1 - \delta)$ for some constant $\delta \in (0, 1/2)$, and $k_n \log k_n \prec -\log b \prec (nk_n)^{1/2}$. Then,*

$$\mathbb{E}_0 \left[P \left\{ \|\Phi\beta - F_0\|_{2,n} > M_n(f_0)^{1/2} \mid Y \right\} \right] = o(1), \quad (10)$$

where

$$M_n(f_0) = \begin{cases} \zeta_n n^{-1}, & \text{if } F_0 \in \mathfrak{L}(\Phi_0), \\ \zeta_n n^{-2\alpha/(1+2\alpha)} \log n, & \text{if } F_0^\top (I - Q_0) F_0 \asymp n, \end{cases}$$

where ζ_n can be any arbitrary sequence that diverges to infinity as n tends to ∞ .

Theorem 3.3 exhibits an adaptive property of the functional horseshoe prior. If the true function is α -smooth, then the posterior contracts around the true function at the near minimax rate of $n^{-\alpha/(2\alpha+1)} \log n$. However, if the true function f_0 belongs to the finite dimensional subspace $\mathfrak{L}(\Phi_0)$, then the posterior contracts around f_0 in the empirical L_2 norm at the parametric $1/\sqrt{n}$ rate. We note that the bound $k_n \log k_n \prec -\log b \prec (nk_n)^{1/2}$ is key to the adaptivity of the posterior, since the strength of the shrinkage towards $\mathfrak{L}(\Phi_0)$ is controlled by b . If $-\log b \prec k_n \log k_n$, then the shrinkage towards $\mathfrak{L}(\Phi_0)$ is too weak to achieve the parametric rate when $F_0 \in \mathfrak{L}(\Phi_0)$. On the other hand, if $-\log b \succ (nk_n)^{1/2}$, the resulting posterior distribution would strongly concentrate around $\mathfrak{L}(\Phi_0)$, and it would fail to attain the optimal nonparametric rate of posterior contraction when $F_0 \notin \mathfrak{L}(\Phi_0)$.

We ignore the subspace of functions such that $\{F \in \mathbb{R}^n : F^\top (I - Q_0) F = o(n), F \notin \mathfrak{L}(\Phi_0)\}$. We only focus on the function space that can be strictly separated from the null space $\mathfrak{L}(\Phi_0)$, although it would be meaningful to illustrate the shrinkage behavior when the regression function f approaches the null space in a sense that $F^\top (I - Q_0) F/n \rightarrow 0$ as $n \rightarrow \infty$.

4 Extensions to Gaussian process priors

Even though the procedure based on the functional horseshoe prior can be interpreted as a partial linear model, its scope of applicability extends to a more general class of nonparametric models. We outline such an extension to Gaussian process (GP) priors below (Neal, 1999; Rasmussen and Williams, 2006):

$$\begin{aligned} F \mid \Sigma, \tau &\sim N(\mathbf{0}, \{\Sigma(\mathbf{x})^{-1} + (I - Q_0)/\tau^2\}^{-1}) \\ \pi(\tau) &\propto \frac{(\tau^2)^{b-1/2}}{(1 + \tau^2)^{(a+b)}} \mathbb{1}_{(0, \infty)}(\tau), \end{aligned} \quad (11)$$

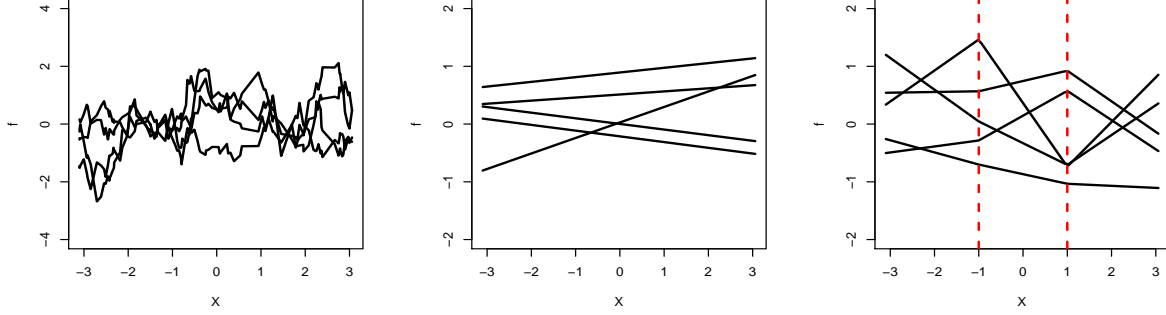


Figure 2: Samples from the classical GP prior (the first column), the GP prior with shrinkage towards linear functions (the second column), and the GP prior with shrinkage towards piecewise linear functions with knots -1 and 1 (the third column).

where $\Sigma(\cdot, \cdot)$ is a positive definite covariance kernel with $\Sigma(\mathbf{x}) = (\Sigma(x_i, x_j))$, and \mathbf{Q}_0 , a and b are defined in (4) and (5). We note that the proposed prior does not define a stochastic process. However, it can be used as a prior on F given the set of locations \mathbf{x} .

To investigate the shrinkage effect of the modified GP prior, we considered two examples of GP priors with the functional shrinkage idea: shrinking towards a class of linear functions and a class of piece-wise linear functions. For shrinking towards linearity, it is straightforward to choose Φ_0 , which is defined in Section 3, as being equivalent to $\{\mathbf{1}, \mathbf{x}\}$. In the same sense, for the shrinkage towards a class of piece-wise linear functions with the knots -1 and 1 , we can consider $\Phi_0 = \{\mathbf{1}, (\mathbf{x} + \mathbf{1})_+, (-\mathbf{x} - \mathbf{1})_+, (\mathbf{x} - \mathbf{1})_+\}$, where $(t)_+ = t$, if $t > 0$ and zero otherwise.

Figure 2 illustrates a comparison between the classical GP prior and the shrinkage version of the GP prior. The covariates were independently generated from a uniform distribution between $-\pi$ and π . The exponential covariance function, i.e., $\Sigma(\mathbf{x})_{i,k} = \exp\{-|x_i - x_k|\}$ for $1 \leq i, k \leq n$, was considered, and we set $a = 1/2$ and $b = n^{-2}$ with sample size $n = 100$. The first plot shows five sample curves generated from the classical GP prior, i.e., $N(\mathbf{0}, \Sigma(\mathbf{x}))$. The second and the third plots display five sample curves from the modified GP prior in (11) with shrinkage towards linear and piece-wise linear functions, respectively. The near parametric forms of the sample paths from the modified GP prior suggest a promising way to shrink GP regression towards simpler parametric classes.

5 Simulation studies

5.1 Univariate examples

In this section, we examine the performance of the functional horseshoe prior on various simulated data sets. We consider three models as follows:

$$\text{i) simple regression model: } Y_i = f(x_i) + \epsilon_i \quad (12)$$

$$\text{ii) varying coefficient model: } Y_i = w_i f(x_i) + \epsilon_i \quad (13)$$

$$\text{iii) density function estimation: } p(Y_i) = \frac{\exp\{f(Y_i)\}}{\int \exp\{f(t)\} dt}, \quad (14)$$

with $\epsilon_i \stackrel{i.i.d}{\sim} N(0, \sigma^2)$ for $i = 1, \dots, n$ in (i) and (ii), and $p(\cdot)$ the density function of Y in (iii). The varying coefficient model (Hastie and Tibshirani, 1993) in (13) reduces to a linear model when the coefficient function

Table 1: Results of univariate examples

True function	Method	$n = 200$	$n = 500$	$n = 1000$
Linear	fHS	0.93 (0.81)	0.44 (0.45)	0.17 (0.17)
	B-spline	3.57 (1.60)	1.54 (0.74)	0.76 (0.38)
Quadratic	fHS	3.63 (1.73)	1.55 (0.74)	0.77 (0.37)
	B-spline	3.59 (1.60)	1.56 (0.74)	0.78 (0.38)
Sine	fHS	3.64 (1.58)	1.50 (0.74)	0.75 (0.36)
	B-spline	3.57 (1.60)	1.53 (0.74)	0.76 (0.38)
Constant	fHS	0.13 (0.15)	0.06 (0.08)	0.03 (0.04)
	B-spline	1.33 (0.63)	0.48 (0.26)	0.25 (0.13)
Quadratic	fHS	1.35 (0.62)	0.51 (0.27)	0.27 (0.13)
	B-spline	1.36 (0.64)	0.51 (0.26)	0.27 (0.13)
Sine	fHS	1.35 (0.63)	0.48 (0.26)	0.25 (0.13)
	B-spline	1.33 (0.63)	0.48 (0.26)	0.25 (0.13)
Normal	fHS	1.34 (1.35)	0.59 (0.52)	0.35 (0.31)
	B-spline	10.30 (5.00)	3.68 (1.42)	1.96 (0.77)
Log-normal	fHS	5.15 (2.70)	3.35 (1.14)	2.91 (0.98)
	B-spline	6.37 (4.21)	3.27 (1.86)	2.83 (1.14)
Mixture	fHS	4.42 (2.18)	1.79 (0.85)	1.04 (0.39)
	B-spline	5.31 (3.61)	1.85 (0.93)	1.04 (0.39)

f is constant, and the density function p is Gaussian when the log-density function f is quadratic in the log-spline model (Kooperberg and Stone, 1991) in (14), motivating the usage of the functional horseshoe prior in these examples to shrink towards the respective parametric alternatives. For each setting, we considered the case corresponding to the relevant parametric model, as well as the parametric model was not adequate.

For (i) and (ii), we generated the covariates independently from a uniform distribution between $-\pi$ and π and set the error variance $\sigma^2 = 1$. For each case (i) - (iii), we considered three parametric choices for f . For case (i), we considered f to be linear, quadratic, and sinusoidal. For case (ii), we considered constant, quadratic and sinusoidal functions. For (iii), we considered normal, log-normal and mixture of normal distributions. For the first two cases, we standardized the true function so as to obtain a signal-to-noise ratio of 1.0.

We used the B-spline basis with $k_n = 8$ in (2) to model the function f in each setting. To shrink the regression function in (12) towards linear subspaces, we set $\Phi_0 = \{\mathbf{1}, \mathbf{x}\}$ in the fHS prior (4). For the varying coefficient model (13), we set $\Phi_0 = \{\mathbf{1}\}$ to shrink f towards constant functions, whence the resulting model reduces to a linear regression model. Finally, we set $\Phi_0 = \{\mathbf{1}, Y, Y^2\}$ to shrink f towards the space of quadratic functions in (14), which results in the density p being shrunk towards the class of Gaussian distributions. We note that the prior for p in (14) is data-dependent. An inverse-gamma prior with parameters $(1/100, 1/100)$ was imposed on σ^2 for the fHS prior in (i) and (ii). In all three examples, we set $b = \exp\{-k_n \log n/2\}$ to satisfy the conditions of Theorem 3.3 and arbitrarily set $a = 1/2$. Although Theorem 3.3 only applies to the regression model (12), the empirical results for these hyperparameter choices are promising for the varying coefficient model and the log-density model as well.

We considered the Jeffrey's prior, $\pi(\beta, \sigma^2) \propto 1/\sigma^2$, on the B-spline coefficients for the simple regression model and the varying coefficient model as a competitor to the functional horseshoe prior. Following Ghosal

et al. (2008), we assigned independent $U(-\pi, \pi)$ priors on the B-spline coefficients, which are known to guarantee the minimax rate of posterior convergence rate for the log-density model. For each prior, we used the posterior mean \hat{f} as a point estimate for f , and report the empirical *Mean Square Error* (MSE), i.e. $\|\hat{f} - f\|_{n,2}^2$.

In Table 1, we report 100 times MSE of the posterior mean estimator and its standard deviation over 100 replicates in estimating the unknown function f for all three models, for sample sizes $n = 200, 500$, and 1000. The first top three rows are for the simple regression model; the second three rows for the varying coefficient model; the last three rows for the density estimation. “Mixture” in the last row indicates a mixture of Gaussian densities as $0.3N(2, 1) + 0.7N(-1, 0.5)$. In all three settings, when the true function f belongs to the nominal parametric class, the posterior mean function resulting from the functional horseshoe prior clearly outperforms the B-spline prior. When the true function does not belong to the parametric model, the functional horseshoe prior performs comparably to the B-spline prior.

Figure 3 depicts the point estimate (posterior mean) and pointwise 95% credible bands for the unknown function f for a single data set for each of the three examples when the true function belongs to the parametric class, that is a linear function in (12), a constant function in (13), and a quadratic function in (14). Figure 4 depicts the corresponding estimates when the data generating function does not fall in the assumed parametric class. It is evident from Figure 3 that when the parametric assumptions are met, the fHS prior performs similarly to the parametric model, which empirically corroborates our findings in Theorem 3.3 that the posterior contracts at a near parametric rate when the parametric assumptions are met. It is also evident that the fHS procedure automatically adapts to deviations from the parametric assumptions in Figure 4, again confirming the conclusion of Theorem 3.3 that when the true function is well-separated from the parametric class, the posterior concentrates at a near optimal minimax rate. We reiterate that the same hyperparameters $a = 1/2$ and $b = \exp\{-k_n \log n/2\}$ for the fHS prior were used in the examples in Figure 3 and Figure 4.

5.2 Comparisons to additive models

Our regression examples in the previous subsection involved one predictor variable. In the case of multiple predictors, a popular modeling framework is the class of additive models (Hastie and Tibshirani (1986), Hastie and Tibshirani (1986)), where the unknown function relating p candidate predictors to a univariate response is modeled as the sum of p univariate functions, with the j th function only dependent on the j th predictor. In this section, we apply the fHS prior to additive models and compare results obtained under this prior to several alternative methods. To be consistent with our previous notation, we express additive models as

$$Y = \sum_{j=1}^p F_j + \epsilon, \quad (15)$$

where $F_j = (f_j(x_{1j}), \dots, f_j(x_{nj}))^T$ for $j = 1, \dots, p$ and $\epsilon \sim N(0, \sigma^2 I_n)$. In the specific case where each f_j is linear, we obtain a linear regression model. In general, each component function can be modeled nonparametrically, for example, using the B-spline basis functions as described in the previous section; $f_j(x) = \sum_{l=1}^{k_n} \beta_{jl} \phi_{jl}(x)$ for $j = 1, \dots, p$. However, if there are many candidate predictors, then nonparametrically estimating p functions may be statistically difficult, and in addition, may result in a loss of precision if only a small subset of the variables are significant. With this motivation, we extend the fHS framework to additive models, where we assign independent fHS priors to the f_j ’s with $Q_0 = 0$ in (4) to facilitate shrinkage of each of these functions towards the null function. We use the resulting posterior mean as a point estimate and compare its performance with a host of penalized likelihood estimators.

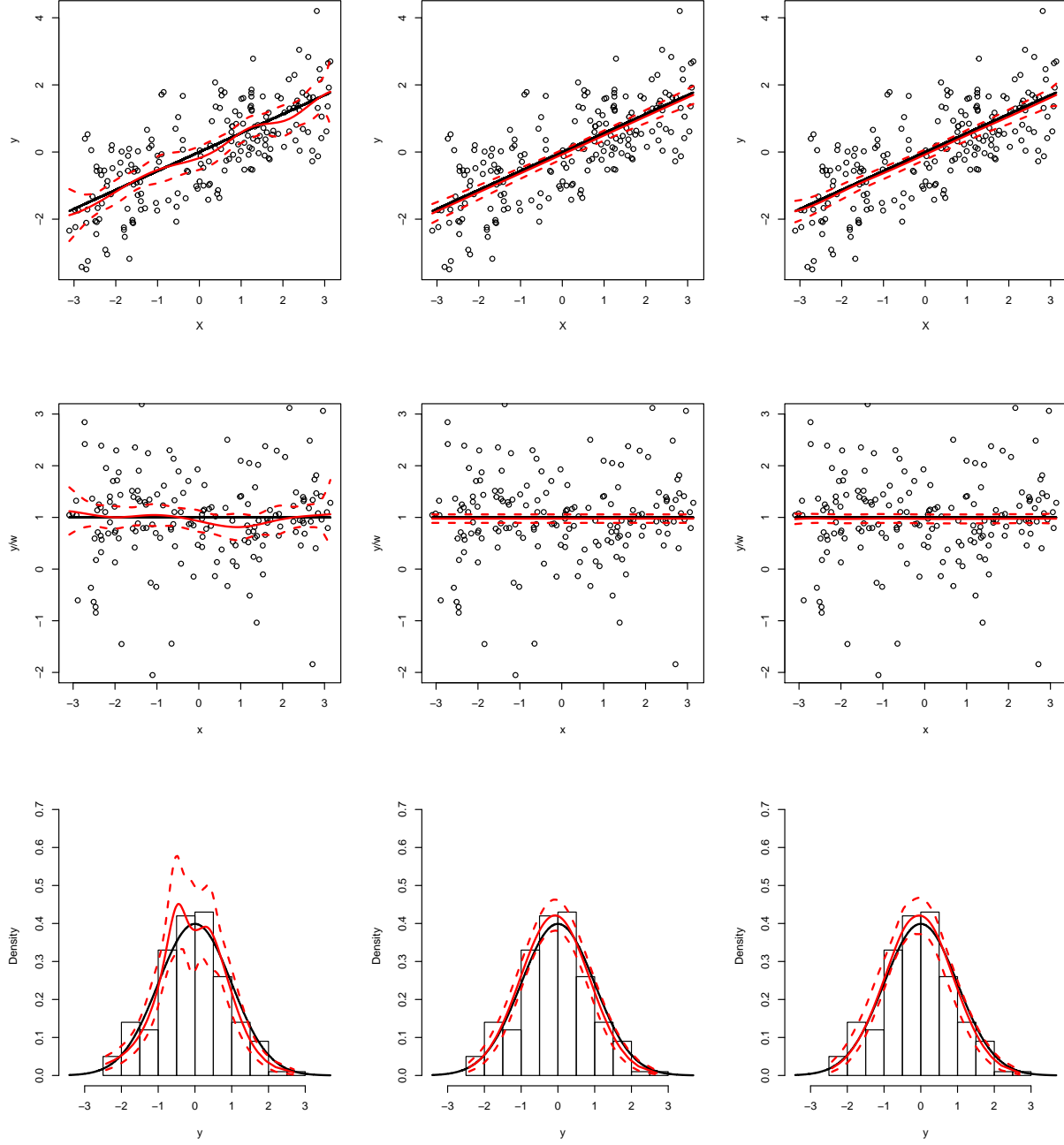


Figure 3: Examples when the underlying true functions are parametric. Posterior mean of each procedure (red solid), its 95% pointwise credible bands (red dashed), and the true function (black solid) from a single example with $n = 200$ for each model. The top row is for the simple regression model; the second row is for the varying coefficient model; the last row is for the density estimation. The Bayesian B-spline procedure, the Bayesian parametric model procedure, and functional horseshoe priors are illustrated in the first, second, and third columns, respectively.

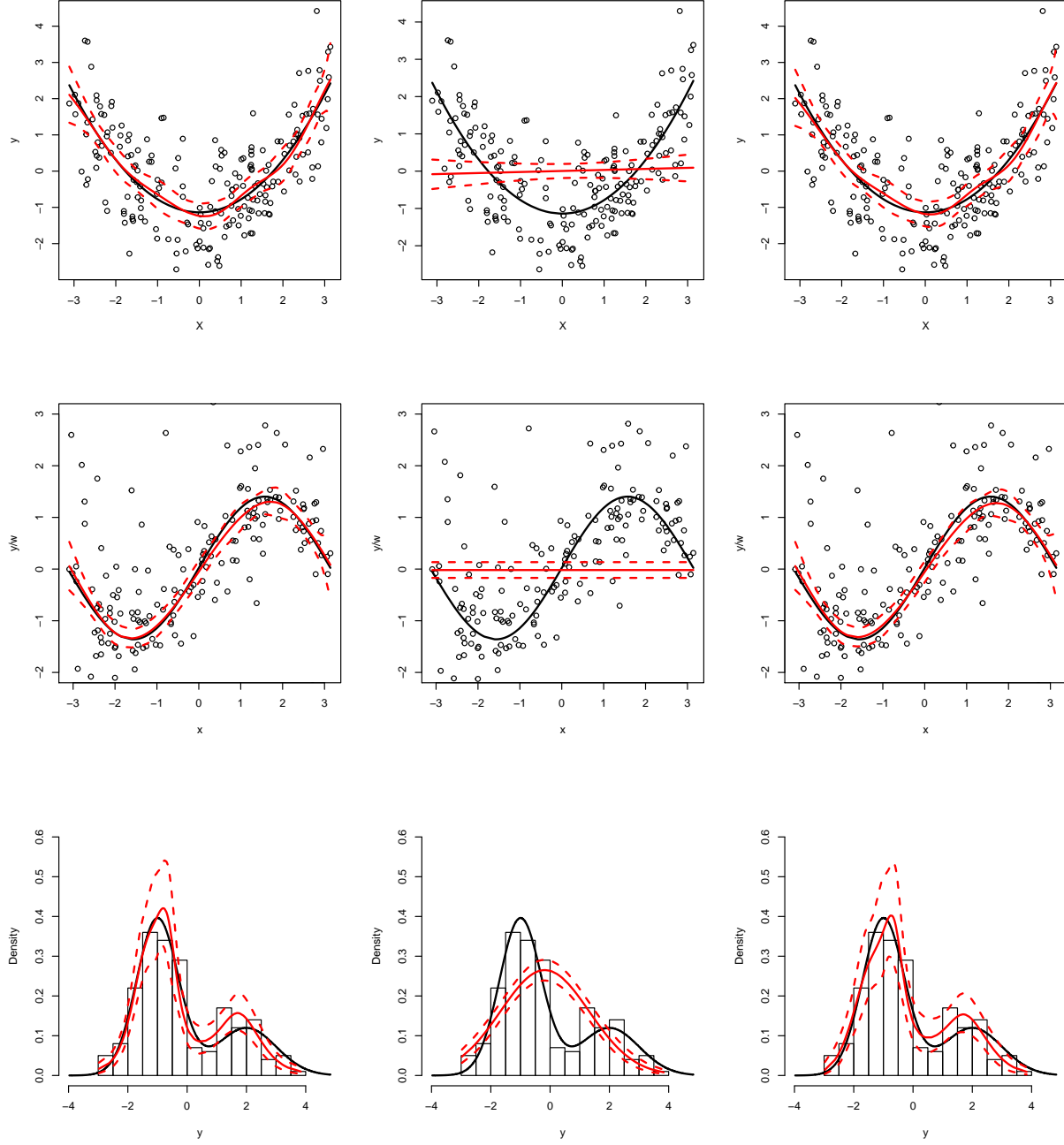


Figure 4: Examples when the underlying true functions are nonparametric. Posterior mean of each procedure (red solid), its 95% pointwise credible bands (red dashed), and the true function (black solid) from a single example with $n = 200$ for each model. The top row is for the simple regression model; the second row is for the varying coefficient model; the last row is for the density estimation. The Bayesian B-spline procedure, the Bayesian parametric model procedure, and functional horseshoe priors are illustrated in the first, second, and third columns, respectively.

For the additive model, Ravikumar et al. (2009) proposed penalized likelihood procedures called *Sparse Additive Models* (SpAM) that combine ideas from model selection and additive nonparametric regression. The penalty term of SpAM can be described as a weighted group Lasso penalty (Yuan and Lin, 2006), and the coefficients for each component function f_j for $j = 1, \dots, p$ are forced to simultaneously shrink towards zero, so that the resulting procedure selects the variables that are associated with the response. Meier et al. (2009) proposed the *High-dimensional Generalized Additive Model* (HGAM) that differs from SpAM in the sense that its penalty term not only imposes shrinkage towards zero, but also regularizes the smoothness of the function. Huang et al. (2010) introduced the two step procedure of adaptive group Lasso (AdapGL) for the additive model, which first estimates the weight of the group penalty, then applies it to the adaptive group lasso penalty. Since the performance of penalized likelihood methods is sensitive to the choice of the tuning parameter, in the simulation studies that follow we considered two criterion for tuning parameter selection: AIC and BIC. R packages `SAM`, `hgam`, and `grplasso` were used to implement SpAM, HGAM, and AdapGL, respectively.

We denote the signal-to-noise ratio as $\text{SNR} = \text{Var}(f(X))/\text{Var}(\epsilon)$, where f is the true underlying function, and we examine the same settings that were considered in Meier et al. (2009) as follows:

Setting 1: ($p = 200$, $\text{SNR} \approx 15$). This is same with Example 1 in Meier et al. (2009), and a similar setting was also considered in Härdle et al. (2012) and Ravikumar et al. (2009). The model is

$$Y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \epsilon_i,$$

where $\epsilon_i \stackrel{i.i.d}{\sim} N(0, 1)$ for $i = 1, \dots, n$, with

$$\begin{aligned} f_1(x) &= -\sin(2x), \quad f_2(x) = x^2 - 25/12, \quad f_3(x) = x, \\ f_4(x) &= \exp\{-x\} - 2/5 \cdot \sinh(5/2). \end{aligned}$$

The covariates are independently generated from a uniform distribution between -2.5 to 2.5 .

Setting 2: ($p = 80$, $\text{SNR} \approx 7.9$). This is equivalent to Example 3 in Meier et al. (2009) and similar with an example in Lin et al. (2006). The model is

$$Y_i = 5f_1(x_{i1}) + 3f_2(x_{i2}) + 4f_3(x_{i3}) + 6f_4(x_{i4}) + \epsilon_i,$$

where $\epsilon_i \stackrel{i.i.d}{\sim} N(0, 1.74)$ for $i = 1, \dots, n$, with

$$\begin{aligned} f_1(x) &= x, \quad f_2(x) = (2x - 1)^2, \quad f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}, \\ f_4(x) &= 0.1 \sin(2\pi x) + 0.2 * \cos(2\pi x) + 0.3 \sin^2(2\pi x) \\ &\quad + 0.4 \cos^3(2\pi x) + 0.5 \sin^3(2\pi x). \end{aligned}$$

The covariate $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$ for $j = 1, \dots, p$ is generated by $\mathbf{x}_j = (W_j + U)/2$, where W_1, \dots, W_p and U are independently simulated from $U(0, 1)$ distributions.

Setting 3 ($p = 60$, $\text{SNR} \approx 11.25$). This is equivalent to Example 4 in Meier et al. (2009), and a similar example was also considered in Lin et al. (2006). The same functions are used and the same process to generate the covariates is considered as in *Setting 2*. The model is

$$\begin{aligned} Y_i &= f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) \\ &\quad + 1.5f_1(x_{i5}) + 1.5f_2(x_{i6}) + 1.5f_3(x_{i7}) + 1.5f_4(x_{i8}) \\ &\quad + 2.5f_1(x_{i9}) + 2.5f_2(x_{i10}) + 2.5f_3(x_{i11}) + 2.5f_4(x_{i12}) + \epsilon_i, \end{aligned}$$

where $\epsilon_i \stackrel{i.i.d}{\sim} N(0, 0.5184)$ for $i = 1, \dots, n$.

To evaluate the estimation performance of the functional horseshoe prior, we report the MSE for each method. To measure the performance of model selection, we considered the proportion of times the true model was selected, as well as the *Matthews correlation coefficient* (MCC; Matthews (1975)), defined as,

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})},$$

where TP, TN, FP, and FN denote the number of true positive, true negatives, false positives, false negatives, respectively. MCC is generally regarded as a balanced measure of the performance of classification methods, which simultaneously takes into account TP, TN, FP, and FN. We note that MCC is bounded by 1, and the closer MCC is to 1, the better the model selection performance is.

For model selection using the fHS prior, we used 95% pointwise credible bands for each component function to exclude component functions whose credible bands uniformly contained the zero function on the entire support of the corresponding covariate. To investigate the performance achieved by the proposed method, we compared it with the partial oracle estimator that refers to the B-spline least squares estimator under the situation where the variables in the true model are given, but the true component functions in the additive model are not provided.

Results from a simulation study to compare these methods are depicted in Figure 5. In all three settings it is clear that the procedure based on the functional horseshoe prior outperforms the penalized likelihood estimators in terms of MSE. In terms of model selection performance, the proposed procedure is also better or at least comparable to that of the competitors. We note that the SpAM procedure with tuning parameter selected by BIC provides comparable model selection performance to the fHS prior in *Setting 1*, yet its MSE is at least 8 times bigger than that of the procedure based on the functional horseshoe prior (note that the reported scale is logarithmic). The results suggest that the fHS prior provides improvement over the penalized likelihood methods in terms of both MSE and model selection performance combined, at least under the considered settings.

6 Real data analysis

In this section, we apply the functional horseshoe prior to two well known data sets: the first concerns ozone levels and the second considers housing prices in Boston. Both data sets are available in the R package `mlbench`. These two data sets have been previously analyzed in various places, including Buja et al. (1989), Breiman (1995), Lin et al. (2006) and Xue (2009). Following the pre-processing step in Xue (2009), we standardized both the response and independent variables prior to our analyses.

We first consider the Boston housing data set that contains the median value of 506 owner-occupied homes in the Boston area, together with several variables that might be associated with the median value. To examine the performance of our method in eliminating extraneous predictors, we add 40 spurious variables generated as i.i.d. standard Gaussian deviates. Using the standard notation for the variable in this data set, we then assumed a model of the following form:

$$\begin{aligned} \text{medv} = & \beta_0 + f_1(\text{crim}) + f_2(\text{indus}) + f_3(\text{nox}) + f_4(\text{rm}) + f_5(\text{age}) + f_6(\text{dis}) + f_7(\text{tax}) \\ & + f_8(\text{ptratio}) + f_9(\text{b}) + f_{10}(\text{lstat}) + \epsilon, \end{aligned}$$

where $\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$. Each component function is modeled by the B-spline bases with $k_n = 8$, and 50 test data points were randomly selected to estimate the out-of-sample prediction error. Five hundreds simulations of each procedure were used to generate the plots in Table 2.

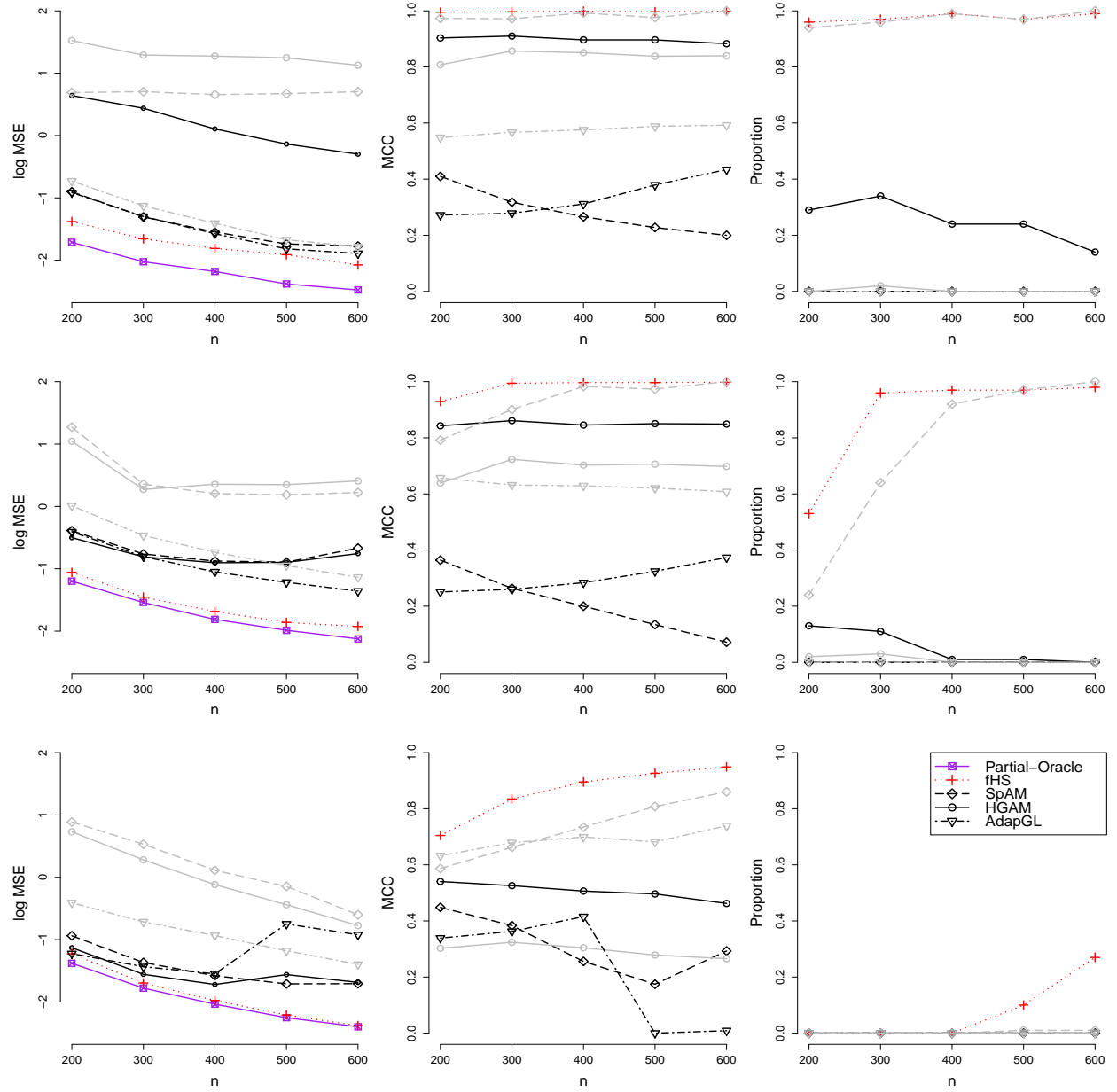


Figure 5: The first column illustrates the logarithm of the MSE of each method; the second column displays the MCC; the third column is the proportion of times the each procedure selected the true model. The top row, the middle row, and the bottom row represent the *Setting 1*, *Setting 2*, and *Setting 3*, respectively. For penalized likelihood methods, AIC (black) and BIC (grey) were used to choose the tuning parameter.

Table 2: Results of real data examples

Data	Method	Test Error	NN	Selected Model
Boston	Full	0.156(0.065)		
	fHS	0.154 (0.067)	0.00	crim, nox, rm, dis, ptratio, lstat
	SpAM(AIC)	0.224(0.072)	21.06	All
	SpAM(BIC)	0.344(0.093)	2.00	crim, nox, rm, dis, ptratio, lstat
	HGAM(AIC)	0.212(0.095)	37.49	All
	HGAM(BIC)	0.222(0.115)	1.06	indus, nox, age, dis, tax, ptratio
	AdaptGL(AIC)	0.579(0.214)	40.00	All
Ozone	Full	0.311(0.085)		
	fHS	0.278 (0.092)	0.02	temp2, gradient
	SpAM(AIC)	0.427(0.156)	20.67	All but height and inv temp
	SpAM(BIC)	0.624(0.213)	0.07	temp1, temp2, gradient
	HGAM(AIC)	0.298(0.109)	23.12	All but gradient
	HGAM(BIC)	0.631(0.260)	0.208	humidity, temp1
	AdaptGL(AIC)	0.359(0.131)	21.91	All but height and inv temp
	AdaptGL(BIC)	0.341(0.142)	2.252	humidity, temp1, temp2, inv height, gradient, visibility

We also modeled the ozone data set using each of the procedures that were applied the housing data. The ozone data consists of the daily maximum one-hour-average ozone readings and nine meteorological variables for 330 days in the Los Angeles basin in 1976. The model applied to these data can be expressed as follows:

$$\begin{aligned}
\text{ozone} = & \beta_0 + f_1(\text{height}) + f_2(\text{wind}) + f_3(\text{humidity}) + f_4(\text{temp1}) + f_5(\text{temp2}) \\
& + f_6(\text{inv height}) + f_7(\text{gradient}) + f_8(\text{inv temp}) + f_9(\text{visibility}) + \epsilon.
\end{aligned}$$

Like the Boston Housing data case, we added 40 spurious variables generated as i.i.d. standard Gaussian deviates. We used B-spline bases with $k_n = 5$ were considered to model the component functions. We performed a cross-validation experiment to assess the predictive performance of the competing methods. In each of 500 simulated data sets, we held out 30 data values as the test set and used the remaining observations to estimate the model. The parameter settings described in Section 5.2 were again used for the functional horseshoe prior. Also, for each training data set we generated 30,000 posterior samples by following the MCMC algorithm described in the Appendix, and only the last 20,000 samples were used in the analysis. We compared the performance of the procedure based on the proposed priors with that of SpAM, HGAM, AdapGL and the classical B-spline estimator was fit without the spurious noise variables. For the penalized likelihood methods, AIC and BIC were used to choose tuning parameters. Table 2 displays the average of test set errors, the average number of selected noise variables, and the most frequently selected model for each method.

In Table 2, “Test Error” refers to the average of empirical L_2 test errors, and “NN” represents the averaged number of selected spurious variables, and “Full” indicates the B-spline least square estimator from the full model without spurious variables. Table 2 shows that for both data sets the procedure based on the functional horseshoe prior achieved the smallest test errors, and it also selected the minimum number of spurious variables. Moreover, even though 40 spurious variables are added to the proposed procedure, its test error was smaller than that of the full estimator that was estimated without the spurious variables. For

both data sets, the model selected by the fHS prior was similar to that chosen by SpAM with BIC. However, the test error of the SpAM procedure was roughly twice that of fHS. More generally, the fHS procedure outperformed all of the other procedures in these examples.

7 Conclusion

We have proposed a class of shrinkage priors which we call the functional horseshoe priors. When appropriate, these priors impose strong shrinkage towards a pre-specified class of functions. The shrinkage term in the prior is new, as it directly allows the nonparametric function shrink towards parametric functions, so it preserves the minimax optimal parametric rate of posterior convergence $n^{-1/2}$ when the true underlying function is parametric, and it also comes within $O(\log n)$ of achieving the minimax nonparametric rate when the true function is strictly separated from the class of parametric functions.

The novel shrinkage term contained in the proposed prior, $F^T(I - Q_0)F$ (i.e., (4)), can be naturally applied to a new class of penalized likelihood methods having a general form expressible as

$$-l(Y | F) + p_\lambda(F^T(I - Q_0)F),$$

where $l(Y | F)$ is the logarithm of a nonparametric likelihood function and p_λ is the penalty term. In contrast to other penalized likelihood, this form of penalty allows shrinkage towards the space spanned by a projection matrix Q_0 , rather than simply a zero function.

References

- Armagan, A., Dunson, D. and Lee, J. (2013) Generalized double Pareto shrinkage. *Statistica Sinica*, **23**, 119–143.
- Bae, K. and Mallick, B. K. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20**, 3423–3430.
- Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015) Dirichlet-Laplace priors for optimal shrinkage. *J. Am. Statist. Ass.*, **110**, 1479–1490.
- Bontemps, D. (2011) Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.*, **39**, 2557–2584.
- Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.
- Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models. *Ann. Statist.*, **17**, 453–510.
- Caron, F. and Doucet, A. (2008) Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*, 88–95. Association for Computing Machinery.
- Carvalho, C., Polson, N. and Scott, J. (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.
- Claeskens, G., Krivobokova, T. and Opsomer, J. D. (2009) Asymptotic properties of penalized spline estimators. *Biometrika*, **96**, 529–544.
- De Boor, C. (2001) *A Practical Guide to Splines*, chap. 9. Springer, Newyork, revised edn.
- Efron, B. and Morris, C. (1973) Stein’s estimation rule and its competitors: an empirical Bayes approach. *J. Am. Statist. Ass.*, **68**, 117–130.
- George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.
- Ghosal, S., Ghosh, J. K., van der Vaart, A. W. et al. (2000) Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.
- Ghosal, S., Lember, J. and van der Vaart, A. (2008) Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, **2**, 63–89.
- Ghosal, S. and van der Vaart, A. (2007) Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, **35**, 192–223.
- Griffin, J. and Brown, P. (2010) Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.
- Hans, C. (2009) Bayesian lasso regression. *Biometrika*, **96**, 835–845.
- Härdle, W. K., Müller, M., Sperlich, S. and Werwatz, A. (2012) *Nonparametric and semiparametric models*. Springer Science & Business Media.
- Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statistical Science*, **1**, 297–318.

- (1993) Varying-coefficient models. *J. R. Statist. Soc. B*, **55**, 757–796.
- Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
- Huang, J., Horowitz, J. L. and Wei, F. (2010) Variable selection in nonparametric additive models. *Ann. Statist.*, **38**, 2282.
- James, W. and Stein, C. (1961) Estimation with quadratic loss. **1**, 361–379.
- Jeffreys, H. (1961) *Theory of Probability*. Clarendon Press, Oxford.
- Johnson, V. E. and Rossell, D. (2012) Bayesian model selection in high-dimensional settings. *J. Am. Statist. Ass.*, **107**, 649–660.
- Kooperberg, C. and Stone, C. J. (1991) A study of logspline density estimation. *Computational Statistics and Data Analysis*, **12**, 327–347.
- Lin, Y., Zhang, H. H. et al. (2006) Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, **34**, 2272–2297.
- Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, **405**, 442–451.
- Meier, L., Van de Geer, S. and Bühlmann, P. (2009) High-dimensional additive modeling. *Ann. Statist.*, **37**, 3779–3821.
- Neal, R. M. (1999) Regression and classification using Gaussian process priors. In *Proceedings of the 6th Valencia World Meeting on Bayesian Statistics*, vol. 6, 475. Oxford University Press.
- (2003) Slice sampling. *Ann. Statist.*, **31**, 705–767.
- Park, T. and Casella, G. (2008) The Bayesian lasso. *J. Am. Statist. Ass.*, **103**, 681–686.
- Polson, N. and Scott, J. (2010) Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*, vol. 9, 501–538. Oxford University Press.
- Polson, N. G. and Scott, J. G. (2012) On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**, 887–902.
- Polson, N. G., Scott, J. G. and Windle, J. (2014) The Bayesian bridge. *J. R. Statist. Soc. B*, **76**, 713–733.
- Rasmussen, C. E. and Williams, C. K. (2006) *Gaussian Process for Machine Learning*. MIT Press, Cambridge.
- Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) *J. R. Statist. Soc. B*, **71**, 1009–1030.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric regression*, vol. 12. Cambridge University Press.
- Tipping, M. (2001) Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, **1**, 211–244.
- Wahba, G. (1990) *Spline models for observational data*. Society for Industrial and Applied Mathematics.

- Xue, L. (2009) Consistent variable selection in additive models. *Statistica Sinica*, **19**, 1281–1296.
- Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, **68**, 49–67.
- Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti*, 233–243. North Holland, Amsterdam.
- Zhou, S., Shen, X., Wolfe, D. et al. (1998) Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, **26**, 1760–1782.

A Proofs of Theorems

Lemma A.1. For arbitrary positive sequences u_n and w_n ,

$$\left(1 - \frac{u_n}{u_n + w_n}\right)^{u_n + w_n} \geq \exp\left\{-u_n + \frac{u_n^2}{2(u_n + w_n)}\right\}. \quad (16)$$

Proof. By Talyor's theorem, there exists $q_n^* \in (0, u_n/(u_n + w_n))$ such that

$$\begin{aligned} \left(1 - \frac{u_n}{u_n + w_n}\right)^{u_n + w_n} &= \exp\left\{(u_n + w_n) \log\left(1 - \frac{u_n}{u_n + w_n}\right)\right\} \\ &= \exp\left\{(u_n + w_n) \left(-\frac{u_n}{u_n + w_n} + \frac{1}{(1 - q_n^*)^2} \frac{u_n^2}{2(u_n + w_n)^2}\right)\right\} \\ &\geq \exp\left\{-u_n + \frac{u_n^2}{2(u_n + w_n)}\right\}. \end{aligned}$$

□

□

Lemma A.2. Suppose W follows a non-central chi-square distribution with m_n degrees of freedom and non-centrality parameter $\lambda_n \geq 0$, i.e, $W \sim \chi_{m_n}^2(\lambda_n)$. Also, let $w_n \rightarrow 0$ and $t_n \rightarrow \infty$ as $n \rightarrow \infty$ and assume that $m_n \prec t_n$. Then,

$$P(W \leq \lambda_n w_n) \leq c_1 \lambda_n^{-1} \exp\{-\lambda_n(1 - w_n)^2\}, \quad (17)$$

and

$$P(W > \lambda_n + t_n) \leq c_2 \left(\frac{t_n}{2m_n}\right)^{m_n/2} \exp\{m_n/2 - t_n/2\} + c_3 \lambda_n^{1/2} t_n^{-1} \exp\left\{-\frac{t_n^2}{32\lambda_n}\right\}, \quad (18)$$

where c_1 , c_2 , and c_3 are some positive constants.

Proof. W can be expressed as $W = \sum_{i=1}^{m_n} \{Z_i + (\lambda_n/m_n)^{1/2}\}^2$, where $Z_i \stackrel{i.i.d}{\sim} N(0, 1)$ for $i = 1, \dots, m_n$. Then, by the fact that $P(Z > a) \leq (2\pi)^{-1/2} a^{-1} \exp\{-a^2/2\}$ for any $a > 0$, we can show that there exist some positive constants c_1 such that

$$\begin{aligned} P(W \leq \lambda_n w_n) &= P\left\{\sum_{i=1}^{m_n} Z_i^2 + 2(\lambda_n/m_n)^{1/2} \sum_{i=1}^{m_n} Z_i + \lambda_n \leq \lambda_n w_n\right\} \\ &\leq P\left\{m_n^{-1/2} \sum_{i=1}^{m_n} Z_i \leq -\lambda_n^{1/2}(1 - w_n)/2\right\} \\ &= P\{|Z_1| \geq \lambda_n^{1/2}(1 - w_n)/2\}/2 \\ &\leq c_1 \lambda_n^{-1} \exp\{-\lambda_n(1 - w_n)^2/2\}, \end{aligned}$$

since Z_1 follows a standard normal distribution.

By using Chernoff's bound and the fact that $P(Z > a) \leq (2\pi)^{-1/2} a^{-1} \exp\{-a^2/2\}$ for any $a > 0$, one can show that

$$\begin{aligned} P(W > \lambda_n + t_n) &= P\left\{\sum_{i=1}^{m_n} Z_i^2 + 2(\lambda_n/m_n)^{1/2} \sum_{i=1}^{m_n} Z_i > t_n\right\} \\ &\leq P\left(\sum_{i=1}^{m_n} Z_i^2 > t_n/2\right) + P\left\{m_n^{-1/2} \sum_{i=1}^{m_n} Z_i > \lambda_n^{-1/2} t_n/4\right\} \\ &\leq c_2 \left(\frac{t_n}{2m_n}\right)^{m_n/2} \exp\{m_n/2 - t_n/2\} + c_3 \lambda_n^{1/2} t_n^{-1} \exp\left\{-\frac{t_n^2}{32\lambda_n}\right\}, \end{aligned}$$

where c_2 and c_3 are some positive constants.

□

□

Proof of Lemma 3.1.

Proof. As discussed in the paragraphs following Lemma 3.1 when $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$, we can generate a new basis $\tilde{\Phi} = [\Phi_0, \Phi_1]$ such that $\Phi_0^\top \Phi_1 = \mathbf{0}$ and $\mathfrak{L}(\Phi) = \mathfrak{L}(\tilde{\Phi})$, which implies $Q_{\tilde{\Phi}} = Q_\Phi$. Then,

$$\begin{aligned}
& \Phi \left(\Phi^T \Phi + \frac{\omega}{1-\omega} \Phi^T (\mathbf{I} - Q_0) \Phi \right)^{-1} \Phi^T \\
&= \tilde{\Phi} \left(\tilde{\Phi}^T \tilde{\Phi} + \frac{\omega}{1-\omega} \tilde{\Phi}^T (\mathbf{I} - Q_0) \tilde{\Phi} \right)^{-1} \tilde{\Phi}^T \\
&= [\Phi_0, \Phi_1] \begin{bmatrix} (\Phi_0^\top \Phi_0)^{-1} & \mathbf{0} \\ \mathbf{0} & (1-\omega)(\Phi_1^\top \Phi_1)^{-1} \end{bmatrix} \begin{bmatrix} \Phi_0^\top \\ \Phi_1^\top \end{bmatrix} \\
&= (1-\omega)Q_{\tilde{\Phi}} + \omega Q_0 \\
&= (1-\omega)Q_\Phi + \omega Q_0.
\end{aligned}$$

□
□

Lemma A.3.

$$n \|Q_0 \Phi \beta - Q_0 Y\|_{n,2}^2 / \sigma^2 \mid Y, \omega \sim \chi_{d_0}^2,$$

and

$$n \|Q_1 \Phi \beta - (1-\omega)Q_1 Y\|_{n,2}^2 / \{(1-\omega)\sigma^2\} \mid Y, \omega \sim \chi_{k_n - d_0}^2.$$

Proof. Recall that

$$\beta \mid Y, \omega \sim \mathcal{N}(\tilde{\beta}_\omega, \tilde{\Sigma}_\omega),$$

where

$$\tilde{\beta}_\omega = \left(\Phi^T \Phi + \frac{\omega}{1-\omega} \Phi^T (\mathbf{I} - Q_0) \Phi \right)^{-1} \Phi^T Y, \quad \tilde{\Sigma}_\omega = \sigma^2 \left(\Phi^T \Phi + \frac{\omega}{1-\omega} \Phi^T (\mathbf{I} - Q_0) \Phi \right)^{-1}.$$

As shown in the proof of Lemma 3.1, $\Phi \left(\Phi^T \Phi + \frac{\omega}{1-\omega} \Phi^T (\mathbf{I} - Q_0) \Phi \right)^{-1} \Phi^T = (1-\omega)Q_\Phi + \omega Q_0$, so

$$\begin{aligned}
\mathbb{E}[Q_0 \Phi \beta \mid Y, \omega] &= Q_0 Y \\
\text{Var}[Q_0 \Phi \beta \mid Y, \omega] &= \sigma^2 Q_0,
\end{aligned}$$

which shows that $n \|Q_0 \Phi \beta - Q_0 Y\|_{n,2}^2 / \sigma^2 \mid Y, \omega \sim \chi_{d_0}^2$.

Similarly,

$$\begin{aligned}
\mathbb{E}[Q_1 \Phi \beta \mid Y, \omega] &= (1-\omega)Q_1 Y \\
\text{Var}[Q_1 \Phi \beta \mid Y, \omega] &= \sigma^2 (1-\omega)Q_1,
\end{aligned}$$

which proves that $n \|Q_1 \Phi \beta - (1-\omega)Q_1 Y\|_{n,2}^2 / \{(1-\omega)\sigma^2\} \mid Y, \omega \sim \chi_{k_n - d_0}^2$. □ □

Proof of Lemma 3.2.

Proof. From Polson and Scott (2012) it follows that

$$\int_0^1 \omega^{A_n-1} (1-\omega)^{B_n-1} \exp\{-H_n \omega\} d\omega = \frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n+B_n)} \exp\{-H_n\} \sum_{m=0}^{\infty} \frac{(A_n)_{(m)}}{(A_n+B_n)_{(m)}} \frac{H_n^m}{m!},$$

where $(a)_{(m)} = a(a+1)\dots(a+m-1)$. We shall show that $\sum_{m=0}^{\infty} \left\{ \frac{(B_n)_{(m)}}{(A_n+B_n)_{(m)}} \frac{H_n^m}{m!} \right\} \geq 1 + Q_n^L$. By using Lemma A.1 and Stirling's approximation, i.e., $m! \asymp m^{m+1/2} \exp\{-m\}$, it follows that

$$\begin{aligned} & \sum_{m=0}^{\infty} \left\{ \frac{(B_n)_{(m)}}{(A_n+B_n)_{(m)}} \frac{H_n^m}{m!} \right\} \\ &= 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + \sum_{m=1}^{\infty} \left[\frac{(B_n+1)_{(m)}}{(A_n+B_n+1)_{(m)}} \frac{H_n^{m+1}}{(m+1)!} \right] \right\} \\ &\geq 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + \sum_{m=1}^{\infty} \left[\frac{(B_n+m)!}{(A_n+B_n+m)!} \frac{H_n^{m+1}}{(m+1)!} \right] \right\} \\ &\geq 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \sum_{m=1}^{\infty} \left[\left(\frac{B_n+m}{A_n+B_n+m} \right)^{A_n+B_n+m+1/2} (B_n+m)^{-A_n} e^{A_n} \frac{H_n^{m+1}}{(m+1)!} \right] \right\} \\ &\geq 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \sum_{m=1}^{T_n} \left[\left(\frac{B_n+1}{A_n+B_n+1} \right)^{1/2} (B_n+m)^{-A_n} \left(\frac{B_n+m}{A_n+B_n+m} \right)^{A_n+B_n+m} e^{A_n} \frac{H_n^{m+1}}{(m+1)!} \right] \right\} \\ &\geq 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \left(\frac{B_n+1}{A_n+B_n+1} \right)^{1/2} (B_n+T_n)^{-A_n} \exp \left\{ \frac{A_n^2}{2(A_n+B_n+T_n)} \right\} \sum_{m=2}^{T_n+1} \frac{H_n^m}{m!} \right\}, \quad (19) \end{aligned}$$

where $T_n = \max\{A_n^2, 3\lceil H_n \rceil\}$, and D is some positive constant.

Since $H_n < (T_n+2) \exp\{1\}$, by using the Stirling's approximation, the term $\sum_{m=2}^{T_n+1} H_n/m!$ in (19) can be expressed as follows:

$$\begin{aligned} \sum_{m=2}^{T_n+1} \frac{H_n^m}{m!} &= \exp\{H_n\} - 1 - H_n - \sum_{m=T_n+2}^{\infty} \frac{H_n^m}{m!} \\ &\leq \exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \sum_{m=T_n+2}^{\infty} \left(\frac{\exp\{1\} H_n}{T_n+2} \right)^m \\ &\leq \exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \end{aligned}$$

Therefore, (19) can be bounded by

$$\begin{aligned} & 1 + \frac{B_n}{A_n+B_n} \left\{ H_n + D \left(\frac{B_n+1}{A_n+B_n+1} \right)^{1/2} (B_n+T_n)^{-A_n} \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \right)_+ \right\} \\ &\geq 1 + \frac{B_n H_n}{A_n+B_n} + \frac{D B_n}{(A_n+B_n)^{3/2}} (B_n+T_n)^{-A_n} \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \right)_+, \end{aligned}$$

where $(\cdot)_+$ denotes the positive hinge function (i.e., for any $t \in \mathbb{R}$, $(t)_+ = t$, if $t > 0$, and $(t)_+ = 0$, otherwise).

Also, since $(B_n+m)!/(A_n+B_n+m)! < 1$ for any positive integer m , it follows that

$$H_n + \sum_{m=1}^{\infty} \left[\frac{(B_n+m)!}{(A_n+B_n+m)!} \frac{H_n^{m+1}}{(m+1)!} \right] \leq \exp\{H_n\},$$

which completes the proof. □

□

Proof of Theorem 3.3. Let β^* denote the projection of the true F_0 on the basis $\{\phi_j\}_{1 \leq j \leq k_n}$, i.e.,

$$\beta^* = \operatorname{argmin}_{\beta \in \mathbb{R}^{k_n}} \|F_0 - \Phi\beta\|_{2,n}. \quad (20)$$

We shall treat β^* as the *pseudo-true* parameter and study the posterior concentration of $\Phi\beta$ in the posterior around $\Phi\beta^*$.

To prove Theorem 3.3, it is sufficient to show that the posterior probability in the equation (10) converges in probability to zero. The quantity in (10) can be decomposed as follows:

$$\begin{aligned} & P \left[\|\Phi\beta - F_0\|_{n,2} > M_n^{1/2} \mid Y \right] \\ & \leq P \left[\|\Phi\beta - \Phi\beta^*\|_{n,2} > M_n^{1/2}/2 \mid Y \right] + \mathbb{1} \left[\|\Phi\beta^* - F_0\|_{n,2} > M_n^{1/2}/2 \right], \end{aligned}$$

where β^* is defined in (20) and $\mathbb{1}(\cdot)$ is the indicator function. The second term on the right-hand side of this expression is always zero when $F_0 \in \mathfrak{L}(\Phi_0)$, since we assume that the column space of Φ_0 is contained in the column space of Φ , and its expectation with respect to the true density is asymptotically zero when $F_0^T(I - Q_0)F_0 \asymp n$ from (9). Therefore, we focus on the first term on the right-hand side. Since $\Phi\beta = Q_1\Phi\beta + Q_0\Phi\beta$, by Lemma 3.1. the first term can be decomposed as

$$\begin{aligned} & P \left[\|\Phi\beta - \Phi\beta^*\|_{n,2} > M_n^{1/2}/2 \mid Y \right] = E_{\omega|Y} \left[P \left(\|\Phi\beta - \Phi\beta^*\|_{n,2} > M_n^{1/2}/2 \mid Y, \omega \right) \right] \\ & \leq E_{\omega|Y} \left[P \left(\|\Phi\beta - \Phi\tilde{\beta}_\omega\|_{n,2} > M_n^{1/2}/4 \mid Y, \omega \right) \right] + E_{\omega|Y} \left[P \left(\|\Phi\tilde{\beta}_\omega - \Phi\beta^*\|_{n,2} > M_n^{1/2}/4 \mid Y, \omega \right) \right] \\ & \leq E_{\omega|Y} \left[P \left(\|Q_1\Phi\beta - (1-\omega)Q_1Y\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega \right) \right] \\ & \quad + E_{\omega|Y} \left[P \left(\|Q_1\Phi\beta^* - (1-\omega)Q_1Y\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega \right) \right] \\ & \quad + E_{\omega|Y} \left[P \left(\|Q_0\Phi\beta - Q_0Y\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega \right) \right] \\ & \quad + \mathbb{1} \left[\|Q_0\Phi\beta^* - Q_0Y\|_{n,2} > M_n^{1/2}/8 \right], \end{aligned}$$

where $\Phi\tilde{\beta}_\omega = (1-\omega)Q_\Phi Y + \omega Q_0 Y = (1-\omega)Q_1 Y + Q_0 Y$.

We denote

$$\begin{aligned} W_1 &= P \left(\|Q_1\Phi\beta - (1-\omega)Q_1Y\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega \right), \\ W_2 &= P \left(\|Q_1\Phi\beta^* - (1-\omega)Q_1Y\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega \right), \\ W_3 &= P \left(\|Q_0\Phi\beta - Q_0Y\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega \right). \end{aligned}$$

The indicator function in the fourth term converges to zero in probability, since $\|Q_0Y - Q_0\Phi\beta^*\|_{2,n}^2$ achieves the parametric optimal rate. To complete the proof we show that the expectations of W_1 , W_2 , and W_3 with respect to the marginal posterior distribution of ω converge to zero in probability.

First consider W_3 . Since $n\|Q_0\Phi\beta - Q_0Y\|_{2,n}^2/\sigma^2 \mid Y, \omega \sim \chi_{d_0}^2$ by Lemma A.3, by using Lemma A.2 it follows that

$$\begin{aligned} E_{\omega|Y} [W_3] &= E_{\omega|Y} \left[P \left\{ \|Q_0\Phi\beta - Q_0Y\|_{2,n} > M_n^{1/2}/8 \mid Y, \omega \right\} \right] \\ &\leq C \left(\frac{nM_n}{64\sigma d_0} \right)^{d_0/2} \exp\{-nM_n/(128\sigma^2)\}, \end{aligned}$$

for some constant C .

The last quantity converges to zero as n tends to ∞ , which implies that $\mathbb{E}_{\omega|Y}[W_3] = o_p(1)$. Now we obtain the bounds on W_1 . By Lemma A.3 $n\|\mathbf{Q}_1\Phi\beta - (1-\omega)\mathbf{Q}_1Y\|_{2,n}^2/\{(1-\omega)\sigma^2\} \mid Y \sim \chi_{k_n-d_0}^2$. By using Lemma A.2, it follows that

$$\begin{aligned} W_1 &\leq \left[\frac{nM_n}{64\sigma^2(k_n-d_0)}(1-\omega)^{-1} \right]^{\frac{k_n-d_0}{2}} \exp \left\{ \frac{k_n-d_0}{2} - \frac{nM_n}{128\sigma^2}(1-\omega)^{-1} \right\} \mathbb{1} \left[\frac{nM_n}{64\sigma^2}(1-\omega)^{-1} > k_n-d_0 \right] \\ &\quad + \mathbb{1} \left[\frac{nM_n}{64\sigma^2}(1-\omega)^{-1} \leq k_n-d_0 \right]. \end{aligned}$$

We denote the two terms in this expression as $W_{1,1}$ and $W_{1,2}$.

By using Lemma 3.2 and defining $\hat{\omega} = (k_n-d_0)/\{nM_n/(64\sigma^2) + k_n-d_0\}$, it follows that

$$\begin{aligned} &\mathbb{E}_{\omega|Y}[W_{1,1}] \\ &= \frac{1}{m(Y)} \left[\frac{nM_n \exp\{1\}}{64\sigma^2(k_n-d_0)} \right]^{\frac{k_n-d_0}{2}} \int_{m_n}^1 \omega^{a+\frac{k_n-d_0}{2}-1} (1-\omega)^{b-\frac{k_n-d_0}{2}-1} \exp \left\{ -\frac{nM_n}{128\sigma^2}(1-\omega)^{-1} - H_n\omega \right\} d\omega \\ &\leq \frac{1}{m(Y)} \left[\frac{nM_n \exp\{1\}}{64\sigma^2(k_n-d_0)} \right]^{\frac{k_n-d_0}{2}} \int_{m_n}^1 \omega^{a-1} (1-\omega)^{b-1} \exp \{-H_n\omega\} d\omega \\ &\quad \times \hat{\omega}^{\frac{k_n-d_0}{2}} (1-\hat{\omega})^{-\frac{k_n-d_0}{2}} \exp \left\{ -\frac{nM_n}{128\sigma^2}(1-\hat{\omega})^{-1} \right\} \\ &= \frac{1}{m(Y)} \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} \int_{m_n}^1 \omega^{a-1} (1-\omega)^{b-1} \exp \{-H_n\omega\} d\omega, \end{aligned} \tag{21}$$

where $m_n = \max[0, 1 - nM_n/\{16\sigma^2(k_n-d_0)\}]$.

Also,

$$\begin{aligned} &\mathbb{E}_{\omega|Y}[W_{1,2}] = P_{\omega|Y} \left[\omega < 1 - \frac{nM_n}{64\sigma^2(k_n-d_0)} \right] \\ &= \frac{1}{m(Y)} \int_0^{1-\frac{nM_n}{64\sigma^2(k_n-d_0)}} \omega^{a+(k_n-d_0)/2-1} (1-\omega)^{b-1} \exp\{-H_n\omega\} d\omega \\ &\leq \frac{1}{m(Y)} \left(\frac{nM_n}{64\sigma^2(k_n-d_0)} \right)^{b-1} \int_0^1 \omega^{a+(k_n-d_0)/2-1} \exp\{-H_n\omega\} d\omega \\ &\leq \left(\frac{nM_n}{64\sigma^2(k_n-d_0)} \right)^{b-1} \frac{\Gamma(a+b+(k_n-d_0)/2)}{\Gamma(a+(k_n-d_0)/2)\Gamma(b)} H_n^{-1} \mathbb{1} \left(1 - \frac{nM_n}{64\sigma^2(k_n-d_0)} \geq 0 \right) \exp\{H_n\} \\ &\quad \times \left[1 + \frac{bH_n}{a+b+(k_n-d_0)/2} + D \frac{b(b+T_n)^{-a-(k_n-d_0)/2}}{(a+b+(k_n-d_0)/2)^{3/2}} \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \right)_+ \right]^{-1}, \end{aligned} \tag{22}$$

where $T_n = \max\{(a+(k_n-d_0)/2)^2, 3\lceil H_n \rceil\}$ and D is some constant.

We now consider two cases: (i) when $F_0 \in \mathfrak{L}(\Phi_0)$ and (ii) when $F_0^T(I - Q_0)F_0 \asymp n$.

Case (i) $F_0 \in \mathfrak{L}(\Phi_0)$:

Recall that in this case $M_n = \zeta_n n^{-1}$ for any arbitrary diverging sequence ζ_n . First, we show that $\mathbb{E}_{\omega|Y}[W_1] \xrightarrow{p} 0$ by proving that $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$ and $\mathbb{E}_{\omega|Y}[W_{1,2}] \xrightarrow{p} 0$.

Applying Lemma 3.2, it follows that (21) is bounded above by

$$\begin{aligned} \mathbb{E}_{\omega|Y} [W_{1,1}] &\leq C \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} \frac{1 + \frac{b}{a+b} \exp\{H_n\}}{1 + \delta_n + u_n \frac{Db}{a+b} (\exp\{H_n\} - 1 - H_n - (T_n + 2)^{-1/2})_+} \\ &\leq C \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} \left(1 + \frac{b}{a+b} \exp\{H_n\} \right), \end{aligned} \quad (23)$$

where $\delta_n = bH_n/(a+b+(k_n-d_0)/2)$ and $u_n = (a+b)(b+T_n)^{-a_n-(k_n-d_0)/2}/(a+b+(k_n-d_0)/2)^{3/2}$ with $T_n = \max\{(a+(k_n-d_0)/2)^2, 3\lceil H_n \rceil\}$, and C and D are some constants.

Since $2H_n \sim \chi_{k_n-d_0}^2$, by Lemma A.2 and defining $q_n = k_n^{-1/2}(\log k_n)^{1/2}(-\log b)^{1/2}$, it follows that

$$P[H_n > k_n q_n/2] \leq \exp\{-ck_n q_n\}, \quad (24)$$

for some constant c . Hence, by the condition that $k_n \log k_n \prec -\log b$, it is clear that $b \exp\{H_n\} = o_p(1)$, which shows that $\mathbb{E}_{\omega|Y} [W_{1,1}] = o_p(1)$.

Similarly, since $\Gamma(b)^{-1} \asymp b$, (22) is bounded by

$$C' b \exp\{H_n\} \left(\frac{nM_n}{64\sigma^2(k_n-d_0)} \right)^{b-1},$$

for some constant C' . By (24), $b \exp\{H_n\} = o_p(1)$, which implies $\mathbb{E}_{\omega|Y} [W_{1,2}] = o_p(1)$.

We next show that $E_{\omega|Y} [W_2]$ converges in probability to zero. Applying Lemma 3.2, it follows that

$$\begin{aligned} \mathbb{E}_{\omega|Y} [W_2] &= \mathbb{E}_{\omega|Y} \left[P[\|(1-\omega)Q_1Y - Q_1\Phi\beta^*\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega] \right] = P_{\omega|Y} \left[\omega < 1 - \left(\frac{nM_n}{64\sigma^2 H_n} \right)^{1/2} \right] \\ &= \frac{1}{m(Y)} \int_0^{1 - \left(\frac{nM_n}{128\sigma^2 H_n} \right)^{1/2}} \omega^{a+(k_n-d_0)/2-1} (1-\omega)^{b-1} \exp\{-H_n\omega\} d\omega \\ &\leq \mathbb{1} \left\{ 1 - \left(\frac{nM_n}{128\sigma^2 H_n} \right)^{1/2} \geq 0 \right\} \frac{1}{m(Y)} \left(\frac{nM_n}{64\sigma^2 H_n} \right)^{(b-1)/2} \int_0^1 \omega^{a+(k_n-d_0)/2-1} \exp\{-H_n\omega\} d\omega \\ &\leq \mathbb{1} \left\{ 1 - \left(\frac{nM_n}{128\sigma^2 H_n} \right)^{1/2} \geq 0 \right\} \frac{\Gamma(a+b+(k_n-d_0)/2)}{\Gamma(b)\Gamma(a+(k_n-d_0)/2)} \left(\frac{nM_n}{128\sigma^2 H_n} \right)^{(b-1)/2} \\ &\quad \times \exp\{H_n\} \left\{ 1 + \delta_n + u_n \frac{Db}{a+b} (\exp\{H_n\} - 1 - H_n - (T_n + 2)^{-1/2})_+ \right\}^{-1} \\ &\leq Cb \left(\frac{nM_n}{128\sigma^2} \right)^{(b-1)/2} H_n^{1/2} \exp\{H_n\}, \end{aligned}$$

where C is some constant, and δ_n and u_n are defined following (23).

From (24), it follows that $b\{nM_n/(128\sigma^2)\}^{(b-1)/2} H_n^{1/2} \exp\{H_n\}$ is bounded by $b\{nM_n/(128\sigma^2)\}^{(b-1)/2} (k_n q_n/2)^{1/2} \exp\{k_n q_n/2\}$ with probability greater than $1 - \exp\{-ck_n q_n\}$ from which it follows that $\mathbb{E}_{\omega|Y} [W_2] = o_p(1)$.

Case (ii) $F_0^T(I - Q_0)F_0 \asymp n$:

Recall that in this case $M_n = \zeta_n n^{-2\alpha/(1+2\alpha)} \log n$ for any arbitrary diverging sequence ζ_n , and δ_n and u_n are defined following (23). From (21) it follows that

$$\begin{aligned} \mathbb{E}_{\omega|Y} [W_{1,1}] &\leq \frac{1}{m(Y)} \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} \int_{m_n}^1 \omega^{a-1} (1-\omega)^{b-1} \exp\{-H_n\omega\} d\omega \\ &\leq C \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} \frac{1 + \frac{b}{a+b} \exp\{H_n\}}{1 + \delta_n + u_n \frac{Db}{a+b} (\exp\{H_n\} - 1 - H_n - (T_n + 2)^{-1/2})_+}, \end{aligned}$$

for some constant C . By Lemma A.2, for any sequence $w_n \rightarrow 0$, H_n is larger than $w_n F_0^T Q_1 F_0 / \sigma^2$ with probability greater than $1 - \exp\{-c F_0^T Q_1 F_0 (1 - w_n)^2 / \sigma^2\}$ for some constant c . Since $F_0^T (I - Q_0) F_0 \asymp n$ implies $F_0^T Q_1 F_0 \asymp n$, the last line in the above display can be expressed as

$$C' \exp \left\{ -\frac{nM_n}{128\sigma^2} (k_n - d_0)^{3/2} (b + T_n)^{(k_n - d_0)/2} \right\} + o_p(1),$$

where $T_n = \max\{(a + (k_n - d_0)/2)^2, 3H_n\}$ and C' is some positive constant. Therefore, to show $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{P} 0$, it is sufficient to prove that $T_n^{(k_n - d_0)/2} \exp\{-nM_n/(128\sigma^2)\} = o_p(1)$. For any $\epsilon > 0$,

$$\begin{aligned} & P \left[T_n^{(k_n - d_0)/2} \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} > \epsilon \right] \\ & \leq P \left[(3H_n)^{(k_n - d_0)/2} \exp \left\{ -\frac{nM_n}{128\sigma^2} \right\} > \epsilon \right] + P[3H_n < (a + (k_n - d_0)/2)^2] \\ & \leq P[\log H_n > \zeta_n \log n] + P[3H_n < (a + (k_n - d_0)/2)^2]. \end{aligned}$$

Since $\zeta_n \rightarrow \infty$ as n tends to ∞ , from (18) in Lemma A.2, it follows that the first term in the above display can be bounded above by $\exp\{-c'(n_n^\zeta - F_0^T Q_1 F_0 / \sigma^2)\}$ for some constant c' . Similarly, from (17) in Lemma A.2, the second term is bounded by $\exp\{-c'' F_0^T Q_1 F_0 / \sigma^2\}$ with some constant c'' , which proves that $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{P} 0$.

Since $nM_n \succ k_n$, the indicator function $\mathbb{1}(1 - nM_n/(64\sigma^2(k_n - d_0)) \geq 0)$ in (22) is zero when n is large enough, which results in $\mathbb{E}_{\omega|Y}[W_{1,2}] \xrightarrow{P} 0$.

The marginal posterior mean of W_2 can be decomposed as

$$\mathbb{E}_{\omega|Y}[W_2] \leq P_{\omega|Y} \left[\|(1 - \omega)Q_1 Y - Q_1 Y\|_{n,2} > \frac{1}{16} M_n^{1/2} \right] + \mathbb{1} \left[\|Q_1 Y - Q_1 \Phi \beta^*\|_{n,2} > \frac{1}{16} M_n^{1/2} \right].$$

Results provided by Zhou et al. (1998) (see equation (9) on page 6) show that the second term in the previous expression is $o_p(1)$. The first term can be expressed as

$$\begin{aligned} & P_{\omega|Y} \left[\omega > \left(\frac{nM_n}{256\sigma^2 H_n} \right)^{1/2} \right] = \frac{1}{m(Y)} \int_{\left(\frac{nM_n}{256\sigma^2 H_n} \right)^{1/2}}^1 \omega^{a+(k_n - d_0)/2 - 1} (1 - \omega)^{b-1} \exp\{-H_n \omega\} d\omega \\ & \leq \frac{1}{m(Y)} \exp \left\{ -H_n^{1/2} (nM_n/(256\sigma^2))^{1/2} \right\} \int_0^1 \omega^{a+(k_n - d_0)/2 - 1} (1 - \omega)^{b-1} d\omega \\ & \leq \left[u_n \exp\{-H_n\} \frac{Db}{a+b} \left(\exp\{H_n\} - 1 - H_n - (T_n + 2)^{-1/2} \right)_+ \right]^{-1} \exp \left\{ -H_n^{1/2} (nM_n/(256\sigma^2))^{1/2} \right\}, \end{aligned}$$

for some positive constant D . Since $H_n/n = O_p(1)$ and $-\log b \prec n^{1/2} k_n^{1/2}$, the above quantity converges in probability to zero, which completes the proof. \square

B Computation Strategy: Slice Sampling

In model (1), the conditional posterior distribution of τ based on the functional horseshoe prior can be expressed as

$$\pi(\tau | Y, \beta) \propto (\tau^2)^{-(k_n - d_0)/2 + b - 1/2} (1 + \tau^2)^{-a-b} \exp\{-\beta^T \Phi^T (I - Q_0) \Phi \beta / (2\sigma^2)\}.$$

By reparameterizing $\eta = 1/\tau^2$, the resulting conditional posterior distribution of η can be expressed as

$$\pi(\eta | Y, \beta) \propto \eta^{a+(k_n - d_0)/2 - 1} \exp\{-\beta^T \Phi^T (I - Q_0) \Phi \beta / (2\sigma^2)\} \frac{1}{(1 + \eta)^{a+b}}.$$

As in Polson et al. (2014), a slice sampling method (Neal, 2003) can be used to sample η from its conditional posterior distribution. The resulting MCMC algorithm is described in Algorithm 1.

Algorithm 1 MCMC algorithm for simple nonparametric regression models

Choose an initial value $\beta^{(0)}$ and $\tau^{(0)}$.

For l in $0 : (L - 1)$

Sample $\beta^{(l+1)}$ from $N(\tilde{\beta}_{\omega^{(l)}}, \sigma^2 \tilde{\Sigma}_{\omega^{(l)}})$, where $\tilde{\beta}_{\omega}$ and $\tilde{\Sigma}_{\omega}$ are defined in (7).

(Slice sampling step) Set $\eta = 1/\tau^{2(l)}$ and $t = (\eta + 1)^{-a-b}$.

Sample $u \sim Unif(0, t)$ and set $t^* = u^{-(a+b)^{-1}} - 1$.

Sample $\eta^* \sim \text{truncated Gamma}(a + (k_n - d_0)/2, \beta^{(l+1)\top} \Phi^\top (\mathbf{I} - \mathbf{Q}_0) \Phi \beta^{(l+1)} / (2\sigma^2))$ on $(0, t^*)$,

Update $\tau^{(l+1)}$ by $\eta^{*-1/2}$.

End.

In the additive model in (15) with a product of the functional horseshoe priors, the conditional posterior distribution of β_j given ω_j and the other coefficients $\beta_{(-j)}$, for $j = 1, \dots, p$, can be expressed as

$$\beta_j \mid \omega_j, \beta_{(-j)}, Y \sim N\left(\tilde{\beta}_{j,\omega}, \sigma^2 \tilde{\Sigma}_{j,\omega}\right),$$

where

$$\tilde{\beta}_{j,\omega} = \tilde{\Sigma}_{j,\omega} \Phi_j^\top r_j, \quad \tilde{\Sigma}_{j,\omega} = (1 - \omega_j) (\Phi_j^\top \Phi_j)^{-1}, \quad r_j = Y - \sum_{l \neq j} \Phi_l \beta_l. \quad (25)$$

It follows that sampling Algorithm 1 can be extended to additive regression models to obtain Algorithm 2 below.

Algorithm 2 MCMC algorithm for additive regression models

Choose an initial value $\beta_j^{(0)}$ and $\tau_j^{(0)}$ for $j = 1, \dots, p$.

For l in $0 : (L - 1)$

For j in $1 : p$

Sample $\beta_j^{(l+1)}$ from $N(\tilde{\beta}_{j,\omega^{(l)}}, \sigma^2 \tilde{\Sigma}_{j,\omega^{(l)}})$, where $\tilde{\beta}_{j,\omega}$ and $\tilde{\Sigma}_{j,\omega}$ are defined in (25).

End.

For j in $1 : p$

(Slice sampling step)

Set $\eta = 1/\tau_j^{2(l)}$ and $t = (\eta + 1)^{-a-b}$.

Sample $u \sim Unif(0, t)$ and set $t^* = u^{-(a+b)^{-1}} - 1$.

Sample $\eta^* \sim \text{truncated Gamma}(a + k_n/2, \beta_j^{(l+1)\top} \Phi_j^\top \Phi_j \beta_j^{(l+1)} / (2\sigma^2))$ on $(0, t^*)$,

Update $\tau_j^{(l+1)}$ by $\eta^{*-1/2}$.

End.

End.
